JLIS.it
Italian Journal of Library and Information Science
Rivista italiana di biblioteconomia, archivistica e scienza dell'informazione

# The research archives in the digital environment: the Sapienza Digital Library project

Maria Guercio, Cecilia Carloni

## 1. New definitions and technical responsibilities for representing and recording the scientific knowledge in digital network environments

One of the most critical problems for research archives (specifically in case of archival resources created in digital form for scientific investigations) is their definition, which is also part of the recognition of their relevance. Except from the small group of specialists involved in documenting, recording, managing and preserving the digital records and heritage of science, the reasons and the solutions for these increasing complexities and the effort to transform the traditional frameworks and tools into efficient and updated proposals have not been fully investigated. In many countries and traditions they seem to be completely ignored or developed at a very low level of service. Publishing the results of the research was in the past, even in the recent past, the best, most common and only means to document and preserve the researchers' scientific work. In the last two decades technological developments have transformed the whole process by making it more complex and

JLIS.it Vol. 6, n. 1 (January 2015)
DOI: 10.4403/jlis.it-10989

challenging. An increasing number of scientists is critically aware of new phenomena (the relevance of their digital archives and data and the increasing risks for their persistency), but not enough engaged in promoting a stronger cooperation with digital heritage curators. A similar lack of coordination (at least if compared to the level of difficulties to solve and to their strategic value) can be still recognized among most professionals (librarians, archivists, data and records curators) involved in the implementation of digital services and in the exploitation of technologies for creating, communicating and preserving the documentary heritage.

Lack of protection and security, lack of controls and responsibilities, fragmentation of data, documents and record aggregations, no standardized models to integrate administrative and scientific resources and identify on a well stated basis public and personal/private resources: all these factors seriously affect the quality of the readability and intelligibility of the scientific data and archives, specifically those created and preserved in the digital environment and with a digital dimension. Of course, these critical aspects imply a number of significant issues to face, like the risk of misalignments, the quality of the documentation to provide for illustrating projects, scholars and research history, for securing data and assessing their provenance and authenticity.

One of the most important aspect to consider (which has been deeply investigated in Sapienza project) is the quality of the information/data representations in a digital dimension. More specifically, we agreed with the basic assumption that the "textual representations" cannot survive without a "knowledge support-systems" as Richard Vines, William P. Hall and Gavan McCarthy clearly illustrate in their 2011 contribution (and we first tried to build such a system in compliance with both the archival principles and the best achievements of digital library communities):

"What makes the knowledge support-systems in the current era fundamentally different from the historical world of print is that the

exchanges of bits and bytes of coded information can now occur more or less at light speed —and that these exchanges can be enacted simultaneously between the varying levels of hierarchy (for example, between individuals and research teams; individuals and teams and a research domain level; or between individuals, teams and research domain level and national or international standards body). Also, at least some components of the cognitive processing function are increasingly being assisted and automated or semi-automated by technology" (Vines, Hall, and McCarthy 2011, 149; see also Dryden 2007).

In order to define more precisely where our efforts of digital archivists have to be addressed in respect with the creation of knowledge support-systems, we believe that this change does not only concern the research intensive networks (as previously stressed). Nowadays an increasing number of sectors are affected by this transformation and requires knowledge support-systems as intended by the authors previously mentioned, and, even more, implies what they called institutional frameworks as a public knowledge space (p. 149). Recent technologies have made possible the availability (even if their preservation is still a challenge in many cases) of an impressive amount of information and data related to the scientific research processes, expressed in various formats other than traditional publications and scripts: structured datasets or raw data, spreadsheets, e-mails, blogs, wikis, videos, but also new forms and types of records relevant to document the research process, its quality and reliability (such as protocols for understanding, agreements, administrative records, audit manuals, research services documentation). To be re-used and maintained this information and the knowledge it represents, must be "widespread or easily discovered and accessed" (p. 159) by the members of the research network. But, first of all, this information must be recorded as contextualised evidence of the research projects (that is not as fragments in the web but as archival sedimentation/accumulation well organized and easy to be explored) and must be maintained in

formats respectful of their nature and available for a qualified retrieval, adequate exploitation, interoperable environments and, last but not least, persistent and authentic preservation.

The main question concerns the fact that shared contexts or, simply, understandable contexts, usable for scientific cooperation among individuals, teams and organizations and/or for monitoring and making available outputs cannot be planned without structured information and interoperable schemas, whatever tools are available for advanced indexing. A correct approach implies also flexible solutions for sharing and exchanging contents of various formats, aware of the changing frameworks. Of course a large use of standards compliant with these aims is the most relevant functional requirement for building e-science open architectures.

Without a common dictionary and a robust conceptual framework, the scientific research heritage, specifically if represented in the new and less controlled forms of datasets or sheets, published on wikis and web portals, is at risk of intelligibility and the efforts made for its protection and exploitation will not be able to face old and new challenges and even less to exploit the technological potentialities and new languages today available. Our authors suggest that

"support-systems are being developed and applied so rapidly that insufficient attention is being paid to the problems of conceptual and terminological confusion at different levels of organisation. There are two sources of such confusion. First, a wide range of personnel from different research domains are designing and enacting standards and schemas that reflect their own narrowly focused professional or social languages. Thus when exchanging information across professional boundaries the schemas used to support data and information exchange can often be incommensurable with other schemas. Second, and perhaps more importantly, insufficient attention is being paid to the challenges associated with harmonising variant schemas that emerge at different levels of hierarchy in the modern research enterprise" (172).

In particular the concept of "public knowledge space" and its main peculiarities, as developed by Gavan McCarthy, seem able to provide the intellectual framework required to transform the very basic and limited functions of existing tools like the digital libraries and the digital archival systems into consistent institutional services needed by the scientific networks (and not only), specifically for exploiting the web space and its potentialities: they imply the "introduction of contextual information management practices; and harmonising variant schemas and standards" (174). As a matter of fact, these principles are at the basis of the archival knowledge, methods and tools, and refer to concepts like provenance and archival description of the contexts which incorporate the spatial and temporal qualities of the content such as its creatorship, the overtime history of its uses, the technological and administrative relationships and the chain of management and custody events and responsibilities.

More specifically, the diversity of contextual information intrinsically related to the scientific resources can and has to be interpreted and manifested at many levels:

- by *contextualizing the institutional space* i.e. by providing the regulations and the policies at its basis, by declaring and making available the information workflows related to the acquisition, ingestion, management and preservation of the data collections and archives, by defining and documenting the standards applied and the related guidelines, by describing the scientific projects and researches involved, by supporting and describing domains dictionaries and taxonomies), but also
- by providing the *contextual information for each producer and its research* and by organizing the research information, data

and archival records with a sufficient degree of *contextual descriptive elements*.

## 2. Contextual information and tools: digital libraries with archival functionality

The archival nature of the research documentation and sources can be difficult to recognize and accept (sometimes also by the institutions of memory themselves and their professional community, included those involved in digital libraries network). The tools developed to make accessible and preserve these outputs are usually and still concentrated to support the creation of digital library and institutional repositories for publications or for isolated items produced and maintained as part of peculiar projects based on their own disciplinary dimension. Educational products, research data and databases, documentary evidence collected in the course of scientific investigations are left to the individual capacities of each researcher (this means that in many cases they are going to be lost) or to Google-like search engines. University departments and research centres are not used to investing their limited resources for a comprehensive collection and preservation of their evidential and scientific memory at the conclusion of a research program/project. The library and archival services are often concentrated on their own traditional heritage, less proactive then required, even when their new digital nature could provide the basis for developing a more advanced service to ensure a qualified and wide access and guarantee a preservation environment (Doorn and Tjalsma 2007 and the other articles published in the same issue).

Specialized and efficient series of tools for identifying, describing, making available and preserving this heritage are required. The compliance with the best archival standards is necessary specifically to ensure the contextual information and the qualified control of any hierarchical structure (not avoidable when high volumes of complex information is implied), but it is also essential to define consistent

workflows for acquisition, appraisal and description, to approve adequate internal policies and to build sustainable services. A flexible and dynamic approach must be in place within a common platform, easy to run by IT departments, but also able to guarantee the research diversity.

The object-oriented approach normally supported by the most common digital libraries, Europeana included, cannot satisfy the complexities of the scientific knowledge representation which implies a hierarchically structured and process-oriented description. In the scientific environment it includes at least three levels of attention:

- *for documenting the research context*: this means many degrees of analysis and information retrieval because "the records of a research enterprise can be situated in an information framework that will enable these records [or data] to be understood not just by the people intimately associated with their creation but by others who have an interest or need" and because "there is a focus on mapping the relationship between information and archival resources created through time and the context within which such resources are created" and this approach is part of the archival methodology"[1],

[1] The authors recognize that "Documenting context is an evolving area of archival practice and this is a good time to start using a different term to cover this area. Context control seems to serve the purpose, and could tentatively be defined as: the process of establishing the preferred form of the name of a records creator, describing the records creator and the functions and activities that produced the records, and showing the relationships among records-creators, and between records creators and records, for use in

- *for supporting architectures and representation models* consistent with the research diversity but also its interoperability,
- *for preserving over time both data/archives and their associated meta-information*, including the relevant documentation of the research processes with specific attention to the evolutionary changes which occur in the digital environment.

## 3. The solutions proposed by the Sapienza Digital Library (SDL): a life-cycle model to govern the research data fragmentation

In line with these functional requirements and based on a standardized approach, Sapienza Digital Library (SDL) has been planned with the goal of identifying, making accessible and preserving in digital form the significant scientific heritage created by the Sapienza researchers. The ambition is to make this material understandable and re-usable both for the scientific community and professionals and for no-academic users. The digital resources are described by the investigators themselves, on the basis of detailed policies and with the support of professionals whose expertise is based both on archival and librarian principles and methodologies.

A special attention is dedicated to the contextualization of the resources (through the definition of flexible partitions and other links), to the provenance information and to the scientific rendition of the research projects and their outcomes. Each collection is described and made accessible by respecting its specific vocabularies and standards. Moreover a validation process is always in place to support a qualified approach. At the same time, the SDL system is

---

archival descriptions" (Vines, Hall, and McCarthy 2011, 178–179; see also Dryden 2007).

implemented with the aim of being easily accessible and respectful of access rights for general users and/or (if required) closed research communities. The main goal (a sort of a mission for the SDL team) is to limit the *anarchy* of the data/records collection and curation, by capturing researchers' attention and interest for a broad and qualified usability under the umbrella of a common institutional infrastructure.

The case studies taken into account (such as collections of archaeological documents or audiovisual archives created in performing arts sectors) have clearly shown need of specialized tools able to support the history of fonds/collections, their internal structure and their logical links and relations with creators and preservers. Among other information types, the SDL carefully describes also the collections partitions and the external relations with other collections, their creators and preservers (Yeo 2012).

In order to better discuss the challenging aims previously summarized, this section is articulated in two areas related to i) the innovative nature of SDL as a *digital library with archival functionality* and ii) the *organization services for researchers and users*. A third section will briefly discuss *open questions and new challenges*.

## 3.1 The nature of SDL

The innovative nature of SDL (if compared with more traditional digital libraries available on the web) is firstly based on the attention paid to the complexities involved in the digitization of multidisciplinary materials and on the aim of ensuring the compliance with methodological specificities of different domains. This goal has implied the definition of a conceptual framework able to provide the respect of contextual information relevant for each scientific domain. The archival principles and methods have been

adopted by the project team[2] to build and implement these contextual dimensions. In particular, the SDL has adopted the archival standards (ISAD, ISAAR and ISDIAH) to describe – at a high level – any kind of collections and archives and their related provenance information (their partitions, their creators and preservers, the scientific projects, methods and techniques behind the digitization processes and/or the creation of digital born materials). The standard MODS has been selected for describing the single digitized or digital born items, according to the Europeana representation model.

A special attention has been dedicated to the definition of the nature of the representation (physical object, reproduction or born digital resource) and, in case of digital representation, to the intellectual relations with the represented physical object and the required level of information mediation. The scientific quality of digital representation of the physical objects and the description and curation of the chain of preservation can be supported only by characterizing the metadata attributes related to the physical object, and those referred to the digital surrogate: the Provided CHO (Provided Cultural Heritage Object) and the related web resource, both represented by the EDM (Europeana Data Model) developed by the Europeana research group.[3]

The digital content aggregators (like Europeana), which do not store the actual digital objects, and of course the related physical resources, which they describe and make available online, have to distinguish clearly the physical (original) object and its digital representation and make their related metadata explicit. Europeana,

[2] The project team includes professionals and researchers of many domains: archivists, librarians, IT scientists, experts for network communication, digital rights and modelling,

[3] *Europeana Data Model Mapping Guidelines*, http://pro.europeana.eu/c/document_library/get_file?uuid=99ce6a74-8e55-4321-917a-65bdff1fe5bc&groupId=51031.

as metadata aggregator, collects metadata on cultural heritage of many European cultural institutions without direct responsibility as far as the originals and the long-term preservation of their digital representations. For this reason, the conceptual analysis has been apparently less complex than in the case of the digital finding aids developed by the institutions of memory. Nevertheless, the question has been clearly defined only recently, thanks to the shift from the initial and too simple representation model called ESE (Europeana semantic elements) to the more complex and more adequate EDM model, which has already been mentioned.

More specifically, the EDM model solves many limits of the previous ESE model: it not only allows for a clear distinction between a physical object and its digital representation, but also supports the capacity of distinguishing the resources and their descriptive metadata and includes basic references for contextual information. The separation of conceptual levels and the new types of information allow the capture and the preservation of information on provenance, the aggregation of different kinds of materials and the implementation of descriptive and administrative metadata according to the nature of the objects (physical, digitized or born digital). These changes have contributed to the solution of relevant semantic ambiguities, with specific reference to terms such as author/creator, dates and place when applied to physical objects. The most critical issues concern the archaeological and museum objects, but the question of a separate description for the original objects owned and their digital surrogate is relevant at a general level for the qualification of the digitization projects and must be specified at the early stage of the collection description with the cooperation of scholars who are expert and responsible for the collections creation and preservation. Thanks to this collaboration the project has been planned on a systematic basis and with more attention – as it happens for the archival description – to the roles and responsibilities involved in the creation process of the collection.

Also the documentation of the project and its implementations have been designed according to the archival methodology which is considered both for ensuring the reliability of the SDL services and for assessing the quality of the repository selected for digital preservation of SDL resources. Record management functions have been put in place for supporting legal evidence and certification processes. They are a crucial requirement for the creator (the SDL service itself, run by Sapienza administration, and imply a record management service) and for preservation repository (at the moment outsourced to Cineca Consortium and developed in compliance with the best international standards and national legislation).

The role of the scientific responsibilities is always recognized and respected. It is defined at collection level and made explicit in the presentation and documentation of each project. Special attention is dedicated to guarantee various degrees of access: an easy access for unskilled common users and advanced functions implemented for requirements of the investigators and educators.

To ensure the interoperability and the consistency of the descriptive information, the adoption of metadata must be based on standards, and more specifically international standards for authority files and thesauri, possibly expressed according to the linked data language.

TGN (Getty Thesaurus of Geographic Names) and Geonames for places, VIAF (Virtual International Authority File) for individuals and corporate bodies, PICO 4.3 (*Thesaurus del Portale della Cultura Italiana*) and Marc Code List for Relators for roles, PICO and the *Nuovo Soggettario di Firenze* are the most important standards adopted for SDL indexing, with the eventual integration of controlled vocabularies based on domains and disciplines. The standardized approach for descriptors, names, dates and places should be able to qualify the access points and to improve the efficiency of indexing and contextualization with reference to capacity of the search engines.

## 3.2. Organizational services

Organizational services have been considered crucial elements for qualifying the project. They are based on the definition of a series of internal policies and specific workflows to increase the efficiency of present and future implementations. They include a flexible system for managing digital rights, which is at present under implementation.

More specifically, the SDL regulation (still in draft) recognizes that the digitized or born digital products are part of the University digital ownership. Among other prescriptions, the rules specify the list of potential digital heritage which could be part of SDL. Among others:

a) the collections of digital objects, the digital archives and the individual digital objects created alongside the Sapienza University scholars research and educational activities, with the exception of scientific publications and other products protected by copyright legislation,

b) the scientific products, specifically those communicated as open access, with respect of the authors intellectual property,

c) the digital collections, archives and objects of scientific and educational relevance whose right for digital communication has been legitimately acquired by the University's bodies,

d) the theses and the final dissertations for courses, masters and PhDs,

The regulation and the following policies (which have been identified as a crucial component of SDL system) state specific workflows and procedures aimed at controlling and qualifying the digitization processes and online communication such as:

- each project must be approved by the SDL management committee,
- the scientific plan and related essential metadata (title, history and collection description) must be defined under the responsibility of a scholar involved in the research project and identified on the basis of his/her knowledge of domain,
- the collection description should include: contexts, responsibilities, creators and preservers history (based on ISAAR standard), definition of partitions and levels, chronological dates, quantity and type of materials (based on ISAD standard), the technological analysis related to the digitization project (the phases of the project, the appraisal motivations, the formats),
- the project must include a detailed information on licensing (which will guide the definition of digitization itself with specific reference to the online accessibility of the collection; the authentication process changes for each degree (open access, Sapienza community or the research group),
- the digital collection can be enriched with other materials (publications, videos, reports) able to integrate and further contextualize the information made available,
- many responsibilities can be identified (technical, managerial, editorial),
- procedures for validation are put in place and include documentation to make possible the overtime verification of the integrity and the authenticity of the digital collections but also to make explicit the selection principles at the basis of the digitization.

The project required a relevant effort to guarantee an efficient portal organization with attention dedicated to the quality of information made available and its effective web surfing.

A road map for preserving born digital and digitized resources and associated metadata is also under implementation. It includes the respect of the main standards (OAIS, METS, ISO 16363), rules for responsibilities and workflows for acquiring and maintaining documentation and for ensuring quality and consistency of processes for submission and for archiving. The preservation is a crucial SDL requirement not only for the continuity and the persistency of the investments, but also because the SDL is already (even partially) a place where to recover born digital materials created and managed by the researchers to support their investigations and to improve the communication for any type of users.

From this perspective, but also for accountability reasons, the pilot phase has already proved the need for an accurate management of the digital resource life cycle, to guarantee the accuracy both of the cataloguing metadata and the identification and organization of the archival records involved in the development of the SDL application. For this reason and for the persistency of the digital library contents and functions, a records management platform (the platform active for University electronic records, called TITULUS) will integrate the SDL by providing services to document general plans and each decision related to individual digitization projects with specific reference to the adoption of policies and to the definition of responsibilities.

## 4. Open questions and new challenges

As previously mentioned, the project is not concluded (we should start with plain service in the course of 2015 but we are already able to provide protection and visibility to our community): the prototype still presents some critical aspects to assess and verify. Among others, the most advanced solutions imply more testing than

expected and must be supported by a large cooperation with other universities and research institutions which operate in the area of cultural heritage. Specific agreements are under definition with the other regional Universities.

An essential condition to verify is the measurement of the effectiveness of this first version which could be transformed into a national framework for the curation of digital heritage created in academic environments. A metric is not easy to define and apply (because of the diversity of the producers and collections involved), but it should include the assessment of political and technical conditions such as:

- the capacity to manage qualified representations and contexts for a significant variety of data/records collections,
- the capacity of capturing the attention of the researchers and persuading them (but not too many at the same time) to commit to SDL the digital outputs of their research both for their "institutional/scientific communities" and for enlarging the SDL users and consumers with less fragmentation and longer lifetime perspectives than those allowed by websites based on single project funding,
- an acceptable and efficient balance of flexibility and standardization for representing scientific knowledge and its documentation.

Of course, the pilot nature of the project is not compatible with conclusive considerations. As mentioned, the project is still under testing and still implies many relevant adjustments and implementations. In any case the enterprise which started three years ago is one of the most ambitious among the existing national initiatives whose aim is to mediate in a digital form the representation of scientific knowledge. As previously mentioned the most difficult task has been the development of an interdisciplinary approach, respectful of the diversities and able to include the best

technical solutions suggested by the professional domains. For archivists and librarians, a key (and not avoidable) question concerns the role that these disciplines will be able to play in such an open environment.

## References

Doorn, Peter, and Heiko Tjalsma. 2007. "Introduction: Archiving Research Data." *Archival Science* 7 (1): 1–20. doi:10.1007/s10502-007-9054-6.

Dryden, J. 2007. "From Authority Control to Context Control." In *Respect for Authority: Authority Control, Context Control an Archival Description*, edited by J. Dryden. Binghamton, NY: Haworth Information Press.

Vines, Richard, William P. Hall, and Gavan McCarthy. 2011. "Textual Representations and Knowledge Support-Systems in Research Intensive Networks." In *Toward Web Semantic: Connecting Knowledge in Academic Research*, edited by B. Cope, M. Kalantzis, and L. Magee, 145–95. Cambridge: Chandos Publishing. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.210.9823&rep=rep1&type=pdf.

Yeo, Geoffrey. 2012. "The Conceptual Fonds and the Physical Collection." *Archivaria* 73. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13384.

MARIA GUERCIO, Sapienza University of Rome, DigiLab. maria.guercio@uniroma1.it.

CECILIA CARLONI, Sapienza University of Rome, DigiLab. cecilia.carloni@uniroma1.it.

---

Guercio, Maria and Cecilia Carloni. "The research archives in the digital environment: the Sapienza Digital Library project". *JLIS.it* 6, 1 (January 2015): Art: #10989. doi: 10.4403/jlis.it-10989

ACKNOWLEDGMENT: Ideas and concepts here analyzed have been partially discussed at the conference held in Paris (7-10 July 2014) and organized by the ICA Section of University and Research Institutions Archives (SUV). Maria Guercio is responsible for paragraphs 1, 2 and 4; Cecilia Carloni is the author of paragraph 3.

ABSTRACT: One of the most critical problems for research archives is their definition. Without a common dictionary and a robust conceptual framework, academic research heritage, specifically if in digital form, is at risk; the efforts made for its preservation and exploitation will not be able to face old and new challenges and even less to exploit technological potentialities and new languages available. The tools developed for making accessible and preserving the academic outputs generally support creation of digital library and repositories for publications or for individual items. Specialized and efficient tools for identifying, describing, making available and preserving this heritage are required. Compliance with acknowledged standards is necessary, but it is also essential to define consistent workflows, approve adequate policies and build sustainable services. The paper will discuss these issues by presenting the Sapienza Digital Library and its goals of identifying, making accessible and preserving significant research heritage in digital form. The ambition is to make it understandable and re-usable both for the scientific community and professionals, and for no-academic users. Digital resources are described by the

investigators themselves on the basis of detailed policies and with the support of professionals from the archival and librarian domains. Special attention is devoted to resources contextualization, to the provenance information and to the presentation of research projects and their outcomes. Collections are described and made accessible taking into consideration their specific domain vocabularies and standards and a validation process is in place to ensure a qualified approach.

KEYWORDS: Data curation, Digital library, Research data, Research data archives, Sapienza Digital Library.

Submitted: 2014-11-15

Accepted: 2014-11-23

Published: 2015-01-15